Automatic Speech Recognition of Quechua Language Using HMM Toolkit *

Rodolfo Zevallos[1,3][0000−0003−0192−7740], Johanna Cordova[2][0000−0001−5950−5271], and Luis Camacho[1][0000−0001−6569−550X]

[1] Pontifical Catholic University of Peru, Av. Universitaria 1801, Lima 15088, Peru
l.camacho@pucp.pe
[2] National Institute of Oriental Languages and Civilisations, Paris, France
johanna.cordova@inalco.fr
[3] National University of Callao, Av. Juan Pablo II 306, Bellavista, Peru
rjzevallos.salazar@gmail.com

Abstract. In this paper, we present the implementation of an Automatic Speech Recognition system (ASR) for southern Quechua language. The software can recognize both continuous speech and isolated words. The ASR was developed using Hidden Markov Model Toolkit (HTK) and the corpus collected by SIMINCHIKKUNARAYKU. A dictionary provides the system with a mapping of vocabulary words to sequences of phonemes; the audio files were processed to extract the speech feature vectors (MFCC) and then, the acoustic model was trained using the MFCC files until its convergence. The paper also describes a detailed architecture of an ASR system developed using HTK library modules and tools. The ASR was tested using the audios recorded by volunteers obtaining a 12.70% word error rate.

Keywords: Quechua · endangered languages · ASR · HTK · HMM.

1 Introduction

The Quechuan languages are amongst the most spoken indigenous languages of America. In Peru, 13.9% of the population have Quechua as their first language, and over 22% reported a Quechuan ethnic background[4]. Although the language is official in the regions with the most speakers, it has little visibility, and despite recent efforts by the Peruvian State to develop services in native languages, the number of speakers is decreasing. One of the reasons why speakers are abandoning their language in favour of Spanish is their perception that Quechua is unsuited to modernity and without economic value. Indeed, historically an oral language, Quechua is still little used in its written form and is almost absent from digital uses, which limits its scope. The moving into the digital scope would thus be a way of supporting States' initiatives for fundamental access to public services in native languages and a crucial step towards the revitalization of these languages.

* This project was supported by CONCYTEC CIENCIACTIVA of the Peruvian government through grant 164-2015-FONDECYT and by PUCP through grant 2017-3-0039/436.

[4] 2017 National Census, https://www.inei.gob.pe/

In this paper, we aims to develop an Automatic Speech Recognizer (ASR) for Southern Quechua, based on Markov's Hidden Models (HMM). This statistical approach is the most widely used method for processing low-resourced languages [13, 14]. The objective of our work is to provide the Quechua speakers with an interface between oral and written communication and to provide the research field with the basic building blocks to develop more complex tools that would broaden the prospects for using Quechua on a daily basis.

2 Background & Related Works

Only a few groups in Latin America and abroad have been working recently for the last years on language technology for Peruvian native languages. The Institute of Andean Amazonian Language and Literature (ILLA)[5] has compiled electronic dictionaries for Quechua, Aymara and Guarani. The group Hinantin[6] at the Universidad Nacional San Antonio Abad del Cusco, has developed, among other things, a text-to-speech system for Cusco Quechua and a Quechua spellchecker plug-in for LibreOffice. Rios [15] developed a language technology toolkit of high quality for Southern Quechua, including the first Quechua dependency treebank.

The SIMINCHIKKUNARAYKU[7] project is led by a community of activists researchers whose vision is that the future of the languages of America depends not only on the preservation efforts, but also on the polyglotism of all citizens, regardless of ethnicity. They have developed: HUQARIQ, a tool for collecting speech corpus; QILLQA, a corpora repository [11]; SIMINCHIK, 97 hours of Southern Quechua speech corpus and the corresponding transcribed text [17]; and 2,500 hours of non annotated corpus.

To our knowledge, there was no Quechua speech dataset before the one compiled by SIMINCHIKKUNARAYKU.

3 The Quechuan languages

Quechua is a linguistic family of South America whose languages are spoken by about 7-8 million people, mainly in Peru, Ecuador and Bolivia [8]. According to Torero's classification, this family is divided into two branches, called Central Quechua (QI) and Peripheral Quechua (QII) [16]. The first is a complex set of varieties currently spoken in the central Andes of Peru. The second is subdivided into three subgroups A, B and C, and covers a geographical area that extends as far north as Colombia and as far south as Argentina. In this work, we will focus on the most widespread and widely spoken variety, Quechua IIC, or Southern Quechua. This subgroup is itself composed of 3 mutually intelligible variants: Ayacucho-Chanca, Cusco-Collao and Santiagueño. As part of the SIMINCHIK project, audio and text corpora were collected for two of these variants : the QUECHUA CHANCA, mainly spoken in the Peruvian department of Ayacucho and surroundings, and the QUECHUA COLLAO, spoken in in the Peruvian

[5] http://www.illa-a.org/wp/
[6] http://hinant.in
[7] https://siminchikkunarayku.pe

departments of Cusco and Puno, and in Bolivia. We will then focus on these two variants in our continued work.

3.1 Quechua phonology

Writing system Quechua is written with a Latin alphabet. In Peru, the graphical system for Quechuan languages is officially determined by a "Quechua pan-dialectal alphabet"[8]. Quechua's graphical and spelling systems are phonological, that is, each letter or digraph corresponds to exactly one phoneme, and the word spelling is regular. Table 1 shows the alphabet for Southern Quechua.

Table 1. Alphabet for Southern Quechua

Vocals	Consonants	Semi-cons.
a, i, u	ch, h, k, l, ll, m, n, ñ, p, q, r, s, sh, t, '	w, y

Phonological features The main phonological feature that differentiates the two variants CHANCA and COLLAO is the occurrence of glottalized and aspirated stops on the occlusive consonants (/ch/, /k/, /p/, /q/, /t/): while this feature is distinctive in Quechua Collao, it is not used in Quechua Chanca. Thus, QUECHUA CHANCA has a total of 15 consonants, most of them voiceless, as shown in Table 2. The glottal and an aspirated version of each occlusive for QUECHUA COLLAO leads to a total of 25 consonants for this variant. Voiced consonants from the Spanish phonemic inventory are also used in the many borrowings.

Table 2. Consonants in the phonemic inventory of QUECHUA CHANCA (IPA)

	Bil	Alv	Pal	Vel	Uvu	Glo
Plosive	p	t	t͡ʃ	k	q	
Nasal	m	n	ɲ			
Fricative		s				h
Lat. Approx.		l	ʎ			
Approximant		ɹ				
Semi-consonants	w		y			

Both quechua dialects are trivocalic : the distinctive vocalic phonemes are /a/, /i/, /u/. Though, in QUECHUA COLLAO, when immediately preceding or following the voiceless uvular stop /q/, /i/ is realized [e] or [ɘ], and /u/ is realized [ɔ] for articulatory reasons. Another phonetic difference between Chanca and Collao is in the realization of the phoneme /q/: while it is pronounced as a stop [q] in Collao, it is realized as a fricative [χ] in Chanca.

[8] Ministerial Resolution 1218-1985-ED

3.2 Quechua morphology

Southern Quechua is agglutinative, with suffixes only, and concatenative: each suffix is added to the previous one without morphological changes in the root or in any of the suffixes. The order of the suffixes varies somewhat from one dialect to another but is stable within a given dialect.

Most roots are monosyllabic or bisyllabic. A syllable is a phonemic unit composed of a nucleus, which is here always a vowel (V) and of margins, which are here consonants (C). Thus, the phonological scheme that describes any Quechua root is the following : (C)V(C)-CV(C) [5].

4 Experiments

4.1 System Description

The Quechua speech recognizer developed is based on a triphone model and was designed and tested using Markov's Hidden Models (HTK) tool.

The Quechua dictionary The statistical approach used in the creation of the Quechua speech recognizer is based on phonemes. As we previously mentioned, the official spelling of Quechua is phonological; it is therefore very simple to build a dictionary of phonemes from Cerrón-Palomino's dictionary [5].

Speech Corpus The corpus used for the construction of the Quechua speech recognizer was the one collected by SIMINCHIKKUNARAYKU. It has 97 hours of audio recorded and transcribed, in the two dialects previously mentioned.

We sampled the total data set, obtaining 8 hours of training data and a 2 hours test corpus, for a total of 16,340 instances. The audio for the training corpus was recorded by 9 collaborators (3 men and 6 women), and the test corpus by 2 collaborators (one male and one female). The table 3 shows the number of collaborators and statements in the training and test data.

Table 3. Number of participants and emissions for the training and testing stage

Stage	Numbers of participants	Utterances
Training	9	11,831
Test	2	4,509
Total	11	16,340

Some audios contained music and Spanish parts. In order to filter the content and split the speech at word level, we used a voice detector based on pyAudioAnalysis [9]. The processed audio files were adjusted to single-channel, 16 kHz sampling, 16-bit precision coding and WAV format.

Records must then be encoded to extract the characteristic vectors. The HCopy tool is responsible for producing coded files (feature vector files).

4.2 System Architecture

The Quechua speech recognizer has a 3-stage architecture: pre-processing, training and testing. In the pre-processing stage, we use the audios and transcripts made by the volunteers to create MFCC files (which contain the data relative to the audio signal) and the main MLF file, which contains all the transcripts. For the training stage, we use the MFCC and MLF files, the Quechua dictionary, the language model and a prototype for the construction of an acoustic model. At the final stage, the performance of the model is evaluated with the test set. Figure 1 shows the general architecture of the system.

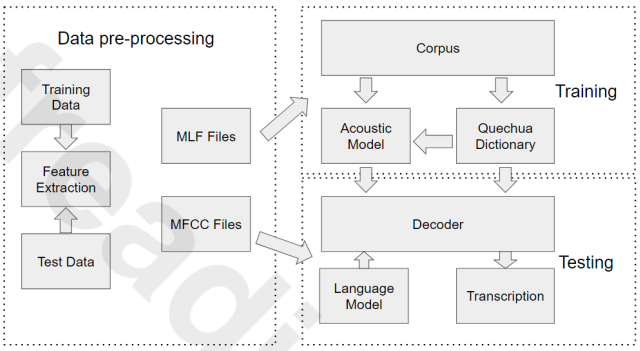

Fig. 1. Quechua Speech Recognition Architecture

4.3 Language Modeling

Given the high presence of rare words in morphologically rich languages such as Quechua, we use singleton pruning rate κ of 0.05 as proposed by [3], to randomly replace only a fraction κ of words that occur only once in the training data with a global UNK symbol.

We built a modified Kneser-Ney model interpolated from 5-gram to word level, getting a perplexity of 298.79. A vocabulary intersection analysis revealed a 63.46% intersection between the validation set and the training set vocabulary. This rather high value of perplexity and low intersection of vocabulary reveals the morphological richness of this language family. A more exhaustive inspection of the vocabulary showed that 64.42% of the words have a frequency of one.

5 Acoustic Models

Different sets of parameters have been tried to check the performance of every arrangement. The base model consists of 5-state HMM in which the first and last states are non-emitting states similar to that presented by Alegre [1]. The model has in each vector of ellipse 39 length, which represents the static vector of (MFCC_0 = 13) plus the

delta coefficients (+13) and the acceleration coefficients (+13). These vectors are extracted from the coded data files. The initialization of means and variances has been done manually for the HMM topology because there are no boot data available (boot data are the data files that have been completely tagged at the data recording stage). After having the HMM prototype file, we must process it using the HCompV tool, which calculates the global averages and variances of the HMM model. The HCompV tool automatically updates the prototype file with new parameters, which will be our starting file for the new estimate. The HERest tool uses the Baum-Welch algorithm to update the states with new parameters (mean, variance and transition probabilities). For each required context-dependent model, the appropriate monophone was cloned and then all of the resulting models were retrained on data using context dependent transcriptions. The next step is to add the word limit. The word limit (pause) is responsible for adding silence between words in continuous speech. Re-estimation is applied again to update model parameters. For the monophonic models, the model is re-estimated using HERES, and for the triphones models, it is re-estimated using HERest and HHEd, which calculate the optimal values of the parameters using Gaussian mixtures. These parameters are re-estimated repeatedly using the training data until it converges.

The experiment follows the procedures mentioned in [12, 7].

6 System Test

The test stage is responsible for generating transcripts of new recordings, for this we use the test data, with a total of 4,509 instances recorded by two collaborators (male and female).

To generate the new transcripts, we must use the HVite tool that is designed to get the recognition files in the Master Label File (MLF) format.

In order to process the test data, the Hvite tool needs the MFCC files of the new recordings, the trained acoustic model, the dictionary and the language model.

7 Results and Discussion

For the comparative analysis between the original transcript and the output sentence of the speech recognizer, which is obtained using the HVite tool, we use the HResults tool, which provides comparative statistics.

Equations (1) show the formula for analyzing the results.

$$\text{Accuracy} = \frac{N - D - S - I}{N} \text{ X } 100\% \tag{1}$$

Where N is the total number of labels in the reference transcripts with correct recognition, D is the number of removal errors, S is the number of substitution errors, and I is the number of insertion errors.

Equation (2) shows the calculation of the word error rate which is a criterion for evaluating speech recognition systems.

$$\text{Word Error Rate} = 100 \text{ - } \%\text{Accuracy} \tag{2}$$

Table 4. Comparative recognition results for different models

Model Set	WER
Monophone Aligned (13 MFCC)	19.60
Monophone Tied (13 MFCC)	14.52
Gaussian Triphone Aligned (13 MFCC)	13.41
Gaussian Triphone Tied (13 MFCC)	12.70

Table 4 gives the recognition performance using monophone aligned (13 MFCC), monophone tied (13 MFCC), Gaussian triphone aligned (13 MFCC), Gaussian and triphone tied (13 MFCC). There are several points to note about these results. Firstly, the overall accuracy is better even for a highly constrained grammar based on the test set itself. This reflects the inherent difficulty of the task. Secondly, the general trend is that increasing the complexity of the models improves performance. Finally, the training data set was deliberately restricted to reduce the overall time needed to train a system. However, even if all of the available data had been used, it seems unlikely that the performance would have exceeded that obtained on the training set. Hence, a closer look was taken at the data. Spontaneous speech is very different from read speech. It contains frequent false starts, repetitions, hesitations, poor articulation, and this problems included missing silence at the beginning and/or end and truncated words at the beginning.

8 Conclusion and future work

This paper presents an automatic speech recognition system for Quechua language based on HTK HMM Toolkit. The results show significant improvements in the WER (reductions from 19.60 to 12.70%) using a Gaussian triphone HMM acoustic model. Although no direct comparison with other published results is possible, it seems that our performance is slightly lower than that proposed by Chuctaya [6] with DTW and KNN. The highly agglutinating nature of language is a challenge for tasks based on word morphology, such as n-gram language modeling, POS labeling, speech recognition, among others. Similar challenges can be found in ASR systems for typologically similar languages such as Turkish [4] or Basque [12]. This system can be used in applications where small vocabulary Quechua speech recognition is required. Moreover, this research work could form the basis for further research in Quechua ASR systems.

An immediate future work is introducing neural networks instead an HMM acoustic model. The new model is to reproduce the work of Graves [10] including extensions made by Amodei [2], that means to introduce neural networks in each stage of the process and use Knowledge Transfer described by Zhao [18] to build a model capable of learning several languages, in this case Basque and Quechua.

References

1. Alegre, F.: Aplicación de rna y hmm a la verificación automática de locutor. IEEE Latin America Transactions **5**(5), 329–337 (2007)

2. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International Conference on Machine Learning. pp. 173–182 (2016)
3. Botha, J.A.: Probabilistic modelling of morphologically rich languages (2015)
4. Carki, K., Geutner, P., Schultz, T.: Turkish lvcsr: towards better speech recognition for agglutinative languages. In: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 3, pp. 1563–1566. IEEE (2000)
5. Cerrón-Palomino, R.: Quechua sureño diccionario unificado quechua-castellano castellano-quechua [unified dictionary of southern quechua, quechua-spanish spanish-quechua]. Lima: Biblioteca Nacional del Perú (1994)
6. Chuctaya, H.F.C., Mercado, R.N.M., Gaona, J.J.G.: Isolated automatic speech recognition of quechua numbers using mfcc, dtw and knn
7. Dua, M., Aggarwal, R., Kadyan, V., Dua, S.: Punjabi automatic speech recognition using htk. International Journal of Computer Science Issues (IJCSI) (4), 359 (2012)
8. Durston, A., Mannheim, B.: Indigenous Languages, Politics, and Authority in Latin America: Historical and Ethnographic Perspectives. University of Notre Dame Pess (2018)
9. Giannakopoulos, T.: pyaudioanalysis: An open-source python library for audio signal analysis. PloS one (12) (2015)
10. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning. pp. 1764–1772 (2014)
11. Melgarejo, N., Camacho, L.: Implementation of a web platform for the preservation of american native languages. In: 2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON). pp. 1–4. IEEE (2018)
12. Odriozola, I., Serrano, L., Hernaez, I., Navas, E.: The ahosr automatic speech recognition system. In: Advances in Speech and Language Technologies for Iberian Languages, pp. 279–288. Springer (2014)
13. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE (2), 257–286 (1989)
14. Rabiner, L.R., Juang, B.H., Rutledge, J.C.: Fundamentals of speech recognition, vol. 14. PTR Prentice Hall Englewood Cliffs (1993)
15. Rios, A.: A basic language technology toolkit for quechua. Procesamiento del Lenguaje Natural (56), 91–94 (2016)
16. Torero, A.: Los dialectos quechuas. Univ. Agraria (1964)
17. Zevallos, R., Camacho, L.: Siminchik: A speech corpus for preservation of southern quechua. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Paris, France (2018)
18. Zhao, Y., Xu, Y.M., Sun, M.J., Xu, X.N., Wang, H., Yang, G.S., Ji, Q.: Cross-language transfer speech recognition using deep learning. In: Control & Automation (ICCA), 11th IEEE International Conference on. pp. 1422–1426. IEEE (2014)